DOCUMENT RESUME

ED 270 475 TM 860 356

AUTHOR Carlson, James E.; Spray, Judith A.
TITLE Analysis of Contingency Tables Involving Multiple-Response Data.
PUB DATE Apr 86
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Cutting Scores; *Data Analysis; Difficulty Level; Error of Measurement; Graphs; High Schools; Language Tests; *Mathematical Models; Mathematics Tests; *Measurement Techniques; Monte Carlo Methods; *Multivariate Analysis; *Regression (Statistics); Research Methodology; Scoring; *Tables (Data); Test Items
IDENTIFIERS ACT Assessment; Bonferroni Procedure; *Contingency Tables; Log Linear Models; Multiple Response Data; Type I Errors

ABSTRACT

This paper discussed methods currently under study for use with multiple-response data. Besides using Bonferroni inequality methods to control type one error rate over a set of inferences involving multiple response data, a recently proposed methodology of plotting the p-values resulting from multiple significance tests was explored. Proficiency categories were based on ACT assessment standard scores on the Mathematics Usage Test and different numbers of categories were compared. Cutting points were chosen in the score distribution so that equal numbers of scores fell in each proficiency category. In addition to loglinear analyses of 40 contingency tables, a logistic regression model in which the actual proficiency scores were included rather than a gross categorization, was employed. Results indicate that including terms for differential discrimination and differential difficulty results in overfitting for many of the items of the ACT Assessment Program Mathematics Usage Test. To examine the performance of methods discussed in this paper for the case in which it is known that there are no true differences, a sample-splitting technique was discussed. Future research suggestions on methods of analysis of multiple-response frequency data included Monte Carlo methods and the use of finite intersection methods. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED270475

Analysis of Contingency Tables Involving Multiple-response Data

James E. Carlson

Judith A. Spray

The American College Testing Program

Paper presented at annual meeting of the American Educational Research Association, San Francisco, April 1986.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J.E. Carlson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 860 356

Analysis of Contingency Tables Involving Multiple-response Data

Introduction

It is unfortunately the case that one need only scan two or three texts with the words "multivariate analysis" in the title to see that there is no concensus among authors as to the body of techniques that comprise this branch of statistical analysis. Some authors (Bock, 1975; Timm, 1975) reserve the term for methods for analyzing multiple random response variables. Others like Marascuilo and Levin (1983) explicitly state that they consider such techniques as multiple regression to be in the domain of multivariate analysis. The difference in opinion that seems to prevail may not be worth an argument but we would like to point out that the very name multivariate implies multiple variates and the term variate is usually reserved for random variables. Hence the authors of this paper exclude multiple regression and other cases of the univariate general linear model from the realm of multivariate analysis.

In the literature on methods for the analysis of contingency tables the distinction between univariate and multivariate methods is less clearly differentiated than is the case of methods for the analysis of continuous variates. Authors of several excellent works on the topic (Goodman, 1970; Bishop, Fienberg & Holland, 1975) include the word multivariate in their titles despite the fact that much of the methodology they present might typically be applied in the case of a single response variate. The problem of definition, as is pointed out by these authors, is that the methodology itself can be applied in the same way whether only one, or more than one, of the variables defining the categories of the contingency table are response variates.

Bock (1975) makes a distinction between single-response and multiple-response data that would appear to be consistent with the distinction between univariate and multivariate techniques in the case of continuous variates. He states:

> If each subject in the sample is assigned, on the basis of his behavioral response, to one of several mutually exclusive categories, the category frequencies thus generated represent single-response data. Thus, the classifying of the response is tantamount to the classifying of the subject (p. 507).

and

> If each subject in the sample makes a number of responses, and each response is assigned to one of several mutually exclusive categories, the resulting frequencies represent multiple-response data ... In general, multiple-response data are more difficult to analyze than single-response data because two stages of sampling are involved -- viz., the sampling of subjects and the sampling of responses within subjects (p. 507).

Bock also points out (p. 552) that multiple response data involves a compound distribution that is difficult to specify and

> The only workable solution at present appears to be the large-sample approach of using the multivariate normal distribution to approximate the distribution of the person parameters after some suitable variate transformation. This is, of course, just the multivariate extension of techniques for the univariate analysis of variance of enumeration data by means of variance-stabilizing transformations ... In the multivariate case, however, variance stabilization is not a sufficient criterion for choosing a transformation because covariances are involved as well (p. 552).

Bock proceeds to present methods for the analysis of multiple-response data based on the multivariate logistic transformation.

It is the need to explain the covariances between response variates as well as their variances that might be thought of as truly differentiating univariate analysis from multivariate analysis. In the usual methods for analyzing contingency tables, of course, this translates to the analysis of interactions (associations) between variables, and, as was pointed out above, there is no distinction between response variates and other variables in such methods.

In our work with test development data we are often in the position of being required to analyze data of the type described by Bock as multiple-response data. In dealing with examinees' responses to more than one item at a time, categorization of individuals' responses is not tantamount to categorization of individuals since each individual has more than one response. Hence we have been examining different ways of handling such data. Methods based on item response theory, and linear structural models for categorical data are obviously applicable but we are looking for alternatives that are less expensive in both sample size requirements and computer time. The primary purpose of this paper is to discuss some methods we are currently studying for use with multiple-response data. The work is in a very early stage and no attempt is made in this paper to reach definitive conclusions. Much more work will need to be done before we can provide suggestions about the potential of the techniques we are examining.

Simultaneous Inference Methods

One way of handling multivariate data is to combine sets of univariate statements into simultaneous statements using such techniques as finite intersection methods or those based on the Bonferroni inequality.

Since the finite intersection techniques require knowledge of the correlations between the test statistics for each variate, and in the cases with which we are dealing these are difficult if not impossible to specify, we have concentrated on Bonferroni inequality methods.

When performing analyses with most modern computer programs it is relatively easy to use the latter technique. One simply computes the p-value that must be attained for significance when k statements are tested at the α level and compares each computed p-value to this criterion. As pointed out by Rosenthal and Rubin (1983), who refer to this technique as using "ensemble-adjusted p values", one can vary the α used with each statement.

p-value Plots

Besides using Bonferroni methods to control type one error rate over a set of inferences involving multiple response data, we have been exploring a recently-proposed methodology of plotting the p-values resulting from multiple significance tests.

Let p_t (t=1,2,...,T) be a set of observed p-values resulting from statistical tests of T hypotheses, H_t. Also let $p_{[i]}$ represent the ith ranked p-value and $N_{[i]}$ the number of observed p-values greater than $p_{[i]}$. Schweder and Spjotvoll (1982) proposed plotting the values of $1 - p_{[i]}$, against the $N_{[i]}$.

The rationale underlying this procedure is that p_t is uniformly distributed on the unit interval if H_t is true for any t. If H_t is not true, on the other hand, p_t will tend to be small, and therefore $(1 - p_t)$ near one. Schweder and Spjotvoll (1982) pointed out that $E(N_{[i]})$ is approximately equal to $T_o(1 - p_{[i]})$ when $p_{[i]}$ is "not too small", and that a plot of $N_{[i]}$ against $(1 - p_{[i]})$ should "for large p indicate a straight line with slope T_o"

(p. 494). That is, the slope of the line is approximately equal to the number of true null hypotheses, T_o. For small values of $p_{[i]}$, $E(N_{[i]})$ will be expected to be greater than $T_o(1 - p_{[i]})$ since false hypotheses are then also counted in $N_{[i]}$. These authors suggested the plots of $N_{[i]}$ against $(1 - p_{[i]})$ should be a straight line for the large p-values toward the left end of the plot, and the number of true null hypotheses could be estimated from the slope of this line. Thus, they argued, when the number of true null hypotheses is unknown, fitting such a line to the plots will provide an estimate of this number. The authors argue that least squares methods should not be applied because the data points in such plots are not independent. Since just the fitting of a single line does not require an independence assumption, presumably they mean that stepwise fitting using least squares and F-statistics should not be used. Fitting a single line by any analytic method would not be possible since doing so would involve a prior choice of how many points to include in the fitting function, tantamount to already having an estimate of T_o. After a line is subjectively fitted to the left-hand portion of the plot, the null hypotheses corresponding to those points in the right-hand portion that deviate above the line are rejected.

Schweder and Spjotvoll (1982) illustrated the use of p-value plots with: tests that all pairwise contrasts in an ANOVA are zero, tests that all k(k-1)/2 correlation coefficients in a k by k matrix are not significantly different from zero, and tests of independence in all possible 2-by-2 subtables from a large contingency table. They also cautioned the reader that positive correlation between the p-values will result in larger sampling variation than would occur for T independent uniformly distributed variates. They provided estimates of sampling variances of the plots for the cases they studied.

The Data Used in the Study

It is often the case in test development that we have data sets involving K contingency tables, in which each table is based on the same N individuals. For example, we may be studying the responses (correct/incorrect) to $N = (N_1 + N_2)$ individuals in two groups, classified into 5 proficiency categories, on 40 test items. Hence we have 40 2-by-2-by-5 contingency tables. Using loglinear models to test hypotheses relating to such contingency tables, the test of the three-way interaction has been referred to in the item-bias literature as a test of non-constant bias (e.g., Marascuilo & Slaughter, 1981). In the context of this paper we use the term differential discrimination, to describe what is being tested by the three way interaction test. We also refer to the test of the group-by-response interaction as a test of differential difficulty. We use these two terms to describe the tests in question because they focus attention on the fact that we are examining differences between the members of two groups on two characteristics, difficulty and discrimination, often used to describe test items. In the primary dataset analyzed for this paper the two groups consisted of a sample used in a "tryout" of items and a sample of examinees who responded to the same items when they were included in a form of the ACT Assessment Mathematics Usage Test. The proficiency categories were based on ACT standard scores in Mathematics Usage and different numbers of categories (three, four, five) were compared. Cutting points were chosen in the score distribution such that, as nearly as possible, equal numbers of scores fell in each proficiency category. Because of the nature of the tryout procedure, the proficiency categories for the tryout sample were based on an accompanying form of the Mathematics Usage Test, independent of the items in the analyses. For the second sample no such scores were available so the

proficiency categories were based on scores that included data from the items being studied. In addition to the loglinear analyses of the 40 contingency tables, a logistic regression model in which the actual proficiency scores were included rather than a gross categorization, was employed.

For comparative puposes some analyses were done in which one sample of examinees on the ACT Assessment Mathematics Usage Test was split in half and analyzed as if these were two different samples. Some analyses using items from the ACT Assessment English Usage test were also performed.

Results

The item tryouts were conducted on separate samples for each item. These samples were of varying sizes averaging near 200. The national test data also involved separate samples for each item, each containing 200 examinees. A complete description of the sampling procedure and table of sample sizes for each item is given in another paper (Spray & Carlson, 1986).

The results for hypotheses of no differential discrimination or differential difficulty between tryout and national administration groups on the Mathematics Usage Test are shown in Table 1 and Figures 1 to 4.

Table 1 shows the observed p-values for the tests of differential discrimination and difficulty between the tryout and national samples. Any given model (e.g., LL3, Loglinear model with 3 proficiency groups; LR, Logistic Regression model) results in 80 statistical tests (40 for differential discrimination and 40 for differential difficulty). Hence the Bonferroni approach to simultaneous inference over all 80 tests using the .05 level of significance results in comparing the p-values to .05/80 = .000625. For each of the three loglinear models only one p-value exceeded this criterion, and in all three cases it was for differential discrimination on

item 28. The logistic regression model agreed with the loglinear analyses with respect to item 28 but also resulted in item 35 being declared to differ in difficulty between the two sets of data. In contrast to the Bonferroni results, if each item was tested individually at the .05 level (that is, each p-value in Table 1 was compared to .05), 8, 7, and 6 items would be declared to have significantly different discriminability, for the three- four- and five-proficiency level models, respectively. Additionally, 7, 3, and 2 items, respectively, would be declared to have significantly different difficulty levels in the two data sets. For the logistic regression model 4 items would show significantly different discrimination and 6 items significantly different difficulty. Of the 40 items, the four models agree that there are no significant differences for 24, that there is a significant discrimination difference for 4, and that there is a significant difficulty difference for 1. The four models are not in agreement for the remaining 11 items.

The p-value plots for the three loglinear models are shown in Figures 1 to 3, and for the logistic regression model in Figure 4. The results for loglinear analysis with 3, 4, and 5 proficiency categories are somewhat linear toward the left with slope about 60 to 65, suggesting about 15 to 20 significant tests out of the 80 performed for each model. The results for the logistic regression model (Figure 4) indicate a lower slope of between 50 and 60 and hence more (20 to 30) significant hypothesis tests. There is, however, deviation from a straight line so this analysis is highly subjective. It is interesting that a lot of this deviation represents smaller-than-expected values of (1 - p) toward the left of the plots, indicative of better-than-expected fit of items to the model. Bishop, Fienberg, and Holland (1975, p. 324) point out that overfitting generally results from fitting models with too many parameters. Following this line of reasoning we would conclude that

including terms for differential discrimination and differential difficulty results in overfitting for many of the items of the ACT Assessment Program Mathematics Usage Test. This effect is even more pronounced in the logistic regression results (Figure 4). The variation among the plots of loglinear analyses with different numbers of proficiency categories emphasizes the well-known fact that chi-square tests of categorized variables are highly dependent on the number of categories chosen and selection of cutting points.

In order to examine the performance of the methods discussed in this paper for the case in which it is known that there are no true differences, a sample-splitting technique was used. Two sets of data involving 20 randomly-selected items from the same Mathematics Usage Test were used and loglinear analyses with 5 proficiency groups were fitted for each item. Tests for differential discrimination and difficulty were performed. Thus there was a total of 40 statistical tests involving each of the two split samples. Using an overall .05 significance level with the Bonferroni method, each p-value reported in Table 2 was compared to .05/40 = .00125 and none was significant. Of the 80 individual statistical tests represented by p-values in Table 2, three were significant at the .05 level. This number agrees fairly closely with the number expected due to chance factors in sampling.

Figure 5 displays the p-value plot for all 80 values shown in Table 2. Since all 80 hypotheses are known to be true, a line of slope 80 should, theoretically, fit the data. It does so fairly well. Comparison of this Figure with Figure 3, also based on loglinear analyses with 5 proficiency groups, is instructive. One can readily reach the conclusion that there are some real differences between the tryout and national samples on the 40 test items used to generate Figure 3.

One analysis of data from the English Usage Test of the ACT Assessment Program was performed in order to illustrate the methods under consideration with a larger (75-item) test. Table 3 shows the p-values from a three-proficiency group loglinear analysis of these data. Of the 150 statistical tests performed, 15 were significant using a Bonferroni approach with an overall alpha level of .05 (hence .05/150 = .00033 was the critical p-value for each test). All 15 significant results were for tests of differential difficulty. If the 75 such tests were considered separately from the tests of discrimination (hence a .05/75 = .00067 critical p-value for each test) an additional 3 items would be declared to have significant differences. Similar analysis of the discrimination differences would not result in additional significant items. If each of the 150 statistical tests were performed separately using the .05 level, 40 items would be declared to differ significantly with respect to difficulty, and 4 items with respect to discrimination.

The p-value plot of the 3-proficiency-group loglinear analysis of the ACT Assessment Program English Usage Test is shown in Figure 6 for all 150 statistical tests. This plot is more similar to those shown by Schweder and Spjotvoll (1982) than any of our other plots. Following those authors' suggestions for subjectively fitting a line to the data, about 110 of the 150 significance tests between tryout and national administration samples on the 75 items would be concluded to represent true null hypotheses and 40 would be declared false. Because the number of items was larger on the English Usage Test than on the Mathematics Usage Test, we decided to examine separate plots of the tests of differential discrimination and differential difficulty. The plot of the discrimination tests (Figure 7) is almost exactly diagonal suggesting all null hypotheses are true. The plot of difficulty p-values

(Figure 8), on the other hand, is linear toward the left with a slope of about 30, suggesting a large number (about 40) of false null hypotheses.

Discussion

In the studies reported in this paper we have only begun to explore the potential of certain methods of analysis for multiple-response frequency data. Any conclusive statements would certainly be premature. We pursued this investigation with a view toward determining whether the cost of more formal studies might be warranted. We think it might be and are considering performing some Monte Carlo investigations. Two important questions providing direction to these studies would be:

1. How much variability is there in the fitted lines in p-value plots under known conditions such as number of statistical tests and number of true null hypotheses?

2. Are there definable relationships between the number of significance tests according to Bonferroni methods and p-value plots, under known conditions?

Another possible area for further research on methods of analysis of multiple-response frequency data is the possibility of using finite intersection methods.

References

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). Discrete multivariate analysis: theory and practice. Cambridge, MA: MIt Press.

Bock, R. D. (1975). Multivariate statistical methods in behavioral research. New York: McGraw Hill.

Fienberg, S. E. (1977). The analysis of cross-classified categorical data. Cambridge, MA: MIT Press.

Goodman, L. A. (1970). The multivariate analysis of qualitative data: Interactions among multiple classifications. Journal of the American Statistical Association, 65, 226-256.

Marascuilo, L. A., & Levin J. R. (1983). Multivariate statistics in the social sciences: A researcher's guide. Monterey: Brooks/Cole.

Marascuilo, L. A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on x^2 statistics. Journal of Educational Measurement, 18, 229-248.

Rosenthal, R., & Rubin, D. B. (1983). Ensemble-adjusted p values. Psychological Bulletin, 94, 540-541.

Schweder, T., & Spjotvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. Biometrika, 69, 493-502.

Spray, J. A., & Carlson, J. E. (1986, April). Comparison of loglinear and logistic regression models for detecting changes on proportions. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Timm, N. H. (1975). Multivariate analysis with applications in education and psychology. Monterey: Brooks/Cole.

Table 1. Mathematics Usage Test p-Values

	Discriminations				Difficulties			
Item	LL3	LL4	LL5	LR	LL3	LL4	LL5	LR
1	0.5540	0.4890	0.5270	0.5430	0.3880	0.9140	0.7200	0.4340
2	0.8230	0.8630	0.8880	0.5430	0.6310	0.6140	0.3100	0.3400
3	0.7520	0.9750	0.6710	0.7150	0.3620	0.6080	0.7100	0.6810
4	0.5840	0.4670	0.6030	0.5360	0.5710	0.2230	0.0340	0.0540
5	0.1380	0.2340	0.0810	0.1800	0.3740	0.0600	0.0730	0.7790
6	0.0960	0.0930	0.0670	0.0780	0.0120	0.1020	0.0660	0.0130
7	0.0040	0.0030	0.0030	0.0040	0.5350	0.0210	0.1020	0.5520
8	0.1160	0.0530	0.0560	0.0810	0.3000	0.4840	0.3310	0.1660
9	0.7190	0.6550	0.6170	0.7860	0.5560	0.1300	0.4790	0.5950
10	0.1280	0.1070	0.1630	0.1040	0.6330	0.4840	0.1810	0.1150
11	0.5320	0.5540	0.6170	0.5750	0.0980	0.3030	0.1780	0.6030
12	0.0970	0.1180	0.0600	0.0970	0.7150	0.4600	0.6170	0.2950
13	0.9750	0.8420	0.9200	0.8280	0.1910	0.8200	0.7520	0.2400
14	0.4130	0.4060	0.3030	0.4170	0.0110	0.3610	0.1250	0.0580
15	0.1120	0.1820	0.2030	0.1680	0.2450	0.5330	0.6150	0.2300
16	0.0160	0.0290	0.0380	0.0140	0.0070	0.0250	0.0610	0.0900
17	0.2820	0.5600	0.4390	0.4800	0.4090	0.1480	0.0710	0.1120
18	0.4800	0.8230	0.6890	0.6470	0.8000	0.8020	0.5460	0.6110
19	0.4430	0.3400	0.2420	0.2930	0.4540	0.0650	0.1980	0.4960
20	0.9200	0.9200	0.6470	0.9030	0.9500	0.4150	0.7610	0.7440
21	0.2400	0.1780	0.1900	0.1580	0.6970	0.6860	0.7430	0.1850
22	0.1250	0.1260	0.2370	0.2030	0.8040	0.9660	0.7270	0.6310
23	0.7910	0.5380	0.7080	0.8110	0.3690	0.4220	0.3100	0.2140
24	0.1730	0.2010	0.2440	0.2560	0.4760	0.1330	0.2650	0.2670
25	0.0190	0.0250	0.0370	0.0520	0.4810	0.8560	0.9950	0.6530
26	0.5900	0.4060	0.6550	0.7090	0.0090	0.0980	0.0850	0.6100
27	0.1340	0.1440	0.0360	0.1120	0.1250	0.5680	0.4350	0.5490
28	0.0000	0.0010	0.0010	0.0000	0.5490	0.7510	0.4980	0.2860
29	0.0270	0.0830	0.0230	0.1180	0.0440	0.0960	0.1800	0.0070
30	0.4430	0.2130	0.1860	0.1820	0.0330	0.1520	0.2890	0.0070
31	0.3250	0.3900	0.3900	0.3350	0.0650	0.1080	0.0550	0.1170
32	0.0360	0.0210	0.2080	0.0840	0.2720	0.3530	0.6170	0.0330
33	0.0180	0.0160	0.0060	0.0080	0.3950	0.3910	0.3560	0.0600
34	0.7910	0.9200	0.6320	0.8270	0.8910	0.1940	0.0630	0.1090
35	0.9750	0.7400	0.8420	0.8740	0.0090	0.0060	0.0030	0.0000
36	0.7400	0.3330	0.6630	0.3760	0.2360	0.4820	0.4280	0.0110
37	0.7520	0.6990	0.4930	0.4670	0.9020	0.8870	0.4900	0.3930
38	0.7080	0.8630	0.7770	0.7730	0.0520	0.4540	0.1500	0.1290
39	0.8880	0.7520	0.8880	0.9500	0.5280	0.7890	0.3510	0.2220
40	0.0420	0.0720	0.2460	0.0560	0.5870	0.6030	0.2740	0.8810

Table 2. Mathematics Usage Test: One Group Split

	First Set		Second Set	
ITEM	Discrim	Difficulty	Discrim	Difficulty
1	0.1172	0.8875	0.1935	0.8065
2	0.5330	1.0000	0.7303	0.5657
3	0.2406	0.1302	0.0826	0.6315
5	0.0879	0.3865	0.5436	0.8625
6	0.6876	0.3198	0.0833	0.0864
7	0.1818	0.9203	0.4779	0.9203
10	0.7743	0.5376	0.6980	0.6892
15	0.7264	0.6391	0.0933	0.8065
16	0.3192	0.6985	0.7277	0.8625
18	0.2288	0.3741	0.3071	0.3681
20	0.7831	0.7642	0.8684	0.0864
25	0.0302	0.6985	0.7370	0.3833
26	0.9067	0.5220	0.5965	0.4709
28	0.5889	0.9203	0.2214	0.2286
29	0.2227	0.2506	0.2788	0.6714
30	0.5809	0.4237	0.1610	1.0000
32	0.0370	0.9203	0.2068	0.0507
33	0.9168	0.2542	0.1083	0.0397
35	0.6598	0.2222	0.7637	0.2146
37	0.7597	0.6892	0.2794	0.8875

Table 3. English Usage Test p-Values

Item	Discrim	Difficulty	Item	Discrim	Difficulty
1	0.6851	0.3985	39	0.2270	0.0008
2	0.3936	0.1035	40	0.7188	0.0003
3	0.9431	0.7694	41	0.4747	0.1383
4	0.6928	0.0099	42	0.4177	0.1879
5	0.6236	0.2818	43	0.8507	0.0091
6	0.7773	0.2043	44	0.7156	0.0117
7	0.0787	0.0172	45	0.5952	0.3184
8	0.6795	0.0001	46	0.7214	0.0033
9	0.0492	0.0000	47	0.3890	0.1354
10	0.1228	0.2101	48	0.8952	0.3602
11	0.9113	0.6174	49	0.9136	0.9943
12	0.1030	0.3864	50	0.7965	0.8523
13	0.2908	0.0141	51	0.7866	0.0089
14	0.7511	0.0063	52	0.7748	0.2302
15	0.2530	0.0000	53	0.7221	0.0011
16	0.6205	0.0012	54	0.6219	0.8529
17	0.4981	0.1103	55	0.3927	0.9918
18	0.7153	0.0042	56	0.5226	0.0059
19	0.2166	0.0025	57	0.7951	0.6820
20	0.4977	0.0215	58	0.7728	0.1697
21	0.3954	0.2600	59	0.7108	0.5677
22	0.2316	0.0573	60	0.3244	0.1658
23	0.3425	0.0000	61	0.0172	0.0259
24	0.2243	0.0000	62	0.5019	0.0000
25	0.3967	0.8113	63	0.4240	0.6503
26	0.4818	0.9233	64	0.0323	0.0212
27	0.9902	0.0494	65	0.2635	0.0000
28	0.8945	0.0000	66	0.6994	0.0104
29	0.9971	0.2293	67	0.4377	0.0397
30	0.4809	0.0288	68	0.2634	0.7094
31	0.3521	0.0000	69	0.0232	0.0001
32	0.0837	0.0000	70	0.4212	0.0000
33	0.2317	0.0001	71	0.3553	0.0204
34	0.5397	0.0117	72	0.2243	0.0626
35	0.5726	0.0544	73	0.6834	0.0312
36	0.1177	0.1886	74	0.8407	0.0103
37	0.4420	0.8675	75	0.1611	0.0000
38	0.6550	0.0000			

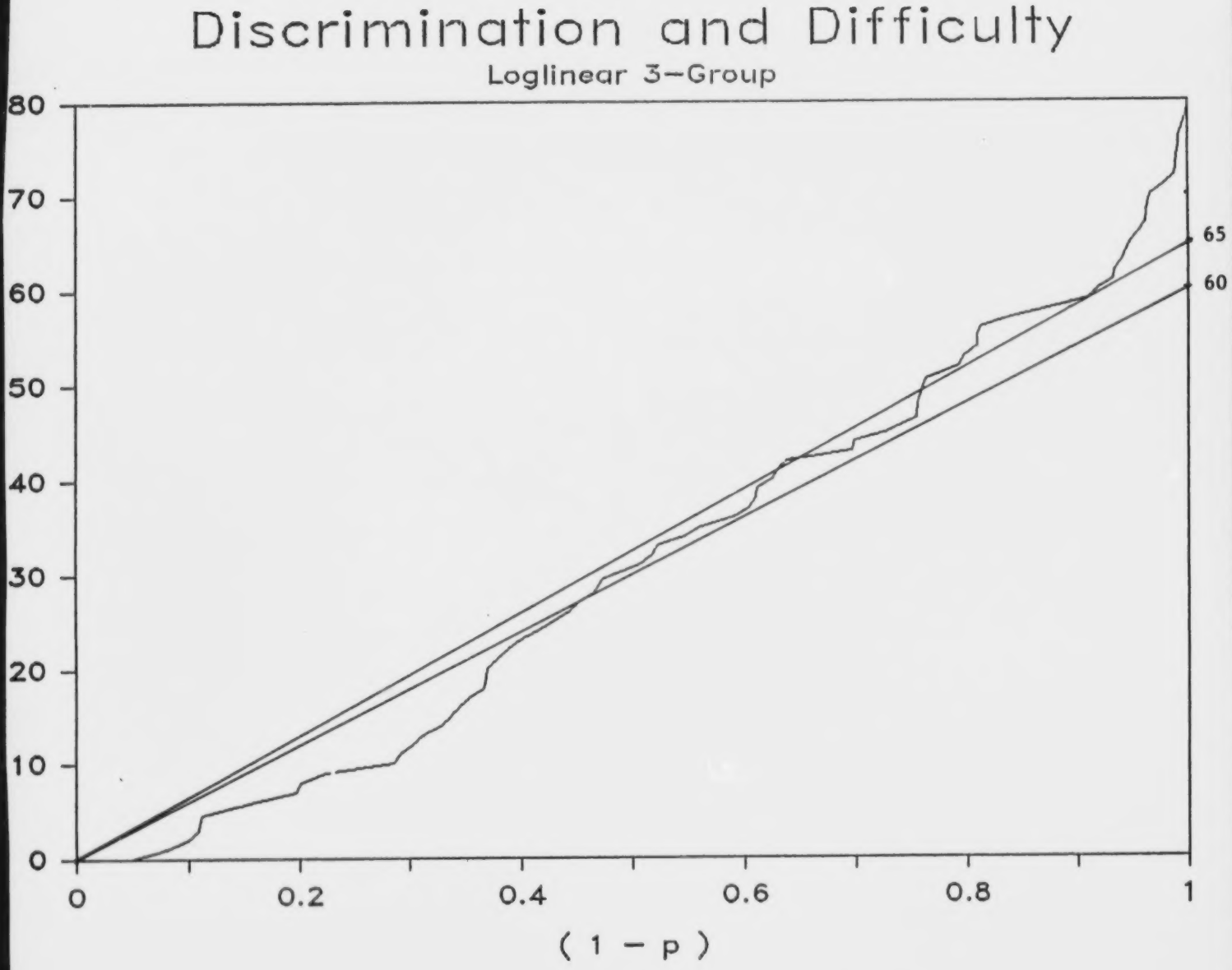
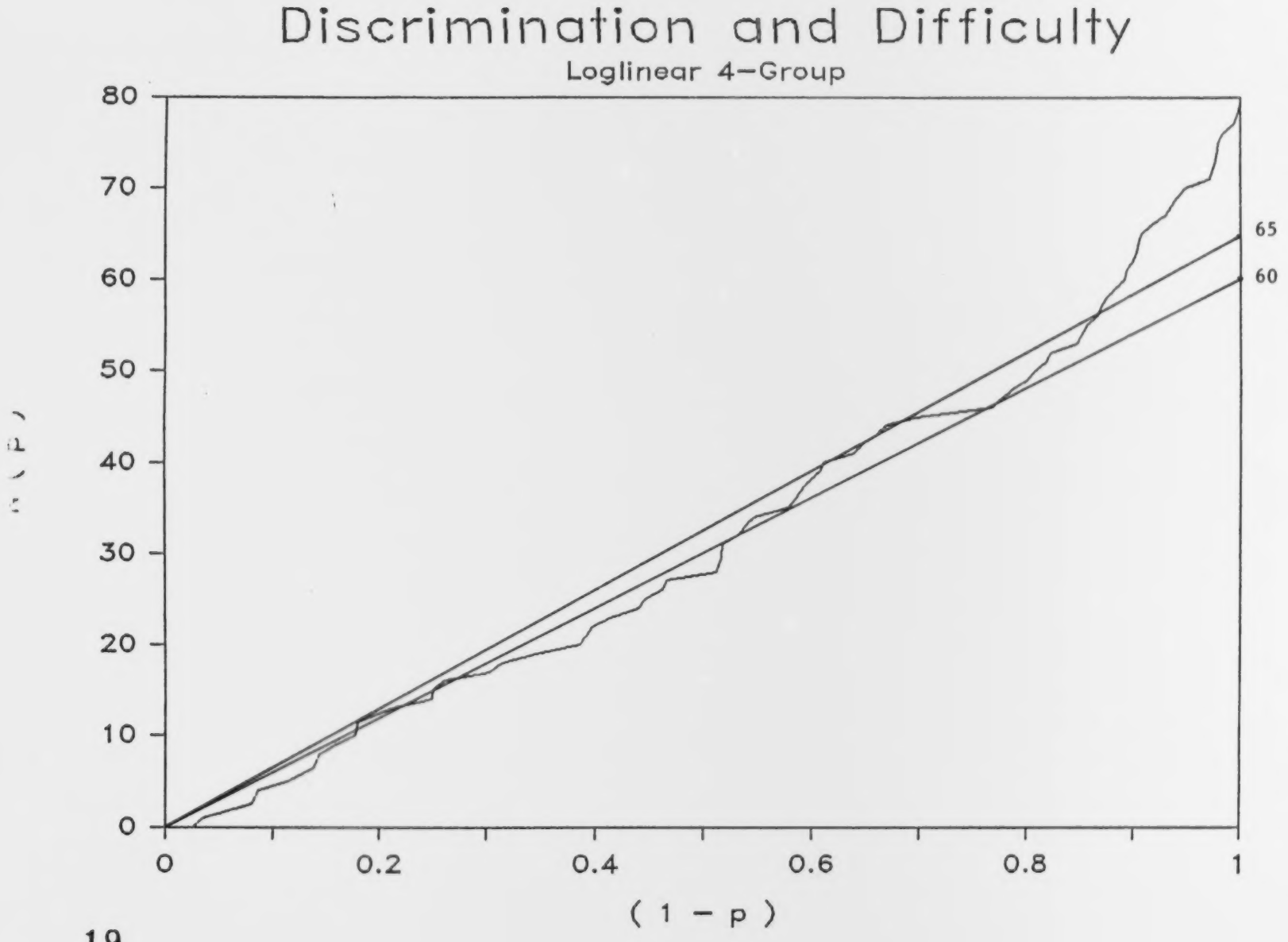

Discrimination and Difficulty
Loglinear 3-Group
80
70
60
50
40
30
20
10
0
0
0.2
0.4
0.6
0.8
1
(1 - p)
65
60

Figure 1



19

Figure 2

20

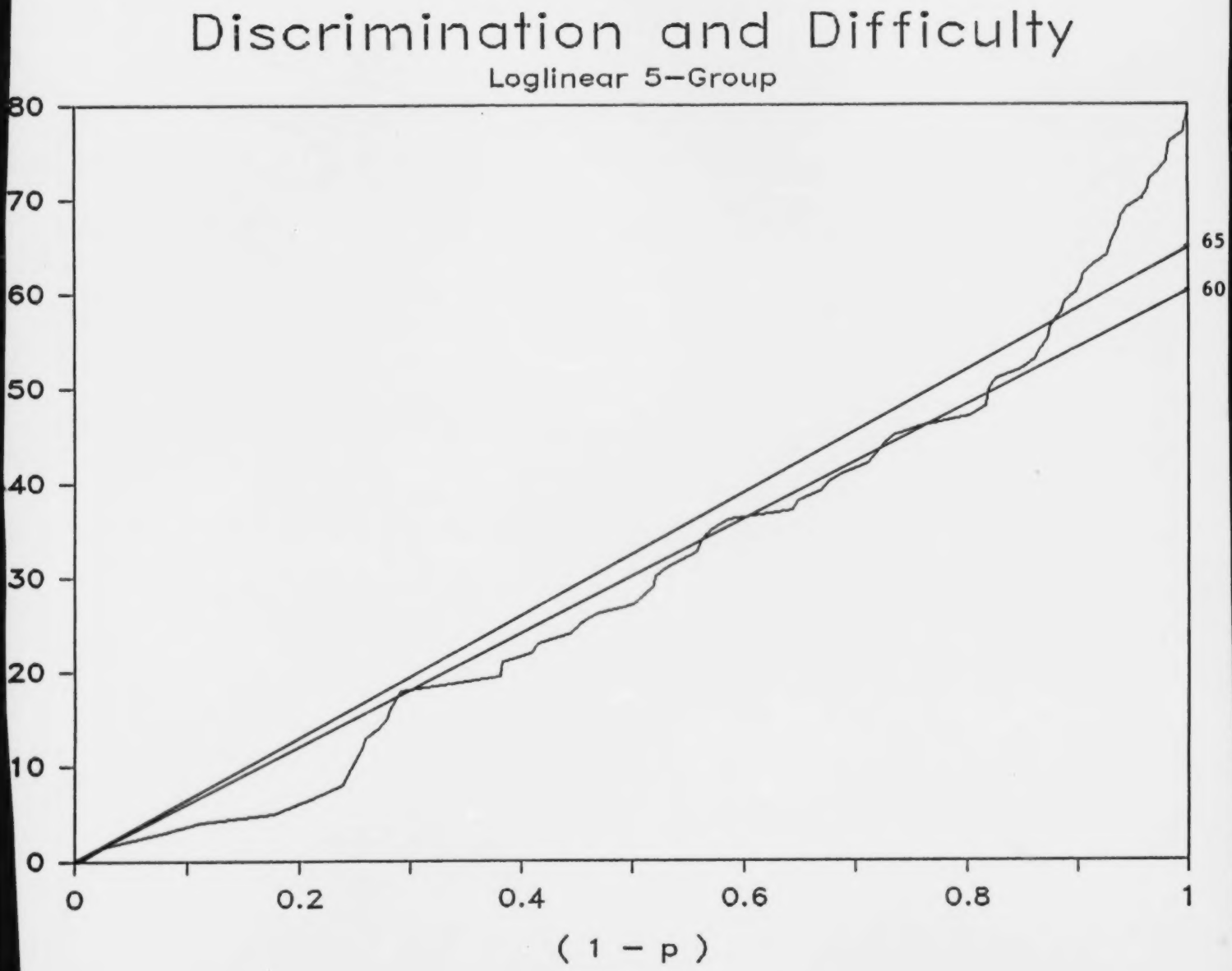

Discrimination and Difficulty
Loglinear 5-Group
80
70
60
50
40
30
20
10
0
0
0.2
0.4
0.6
0.8
1
(1 - p)
65
60

Figure 3

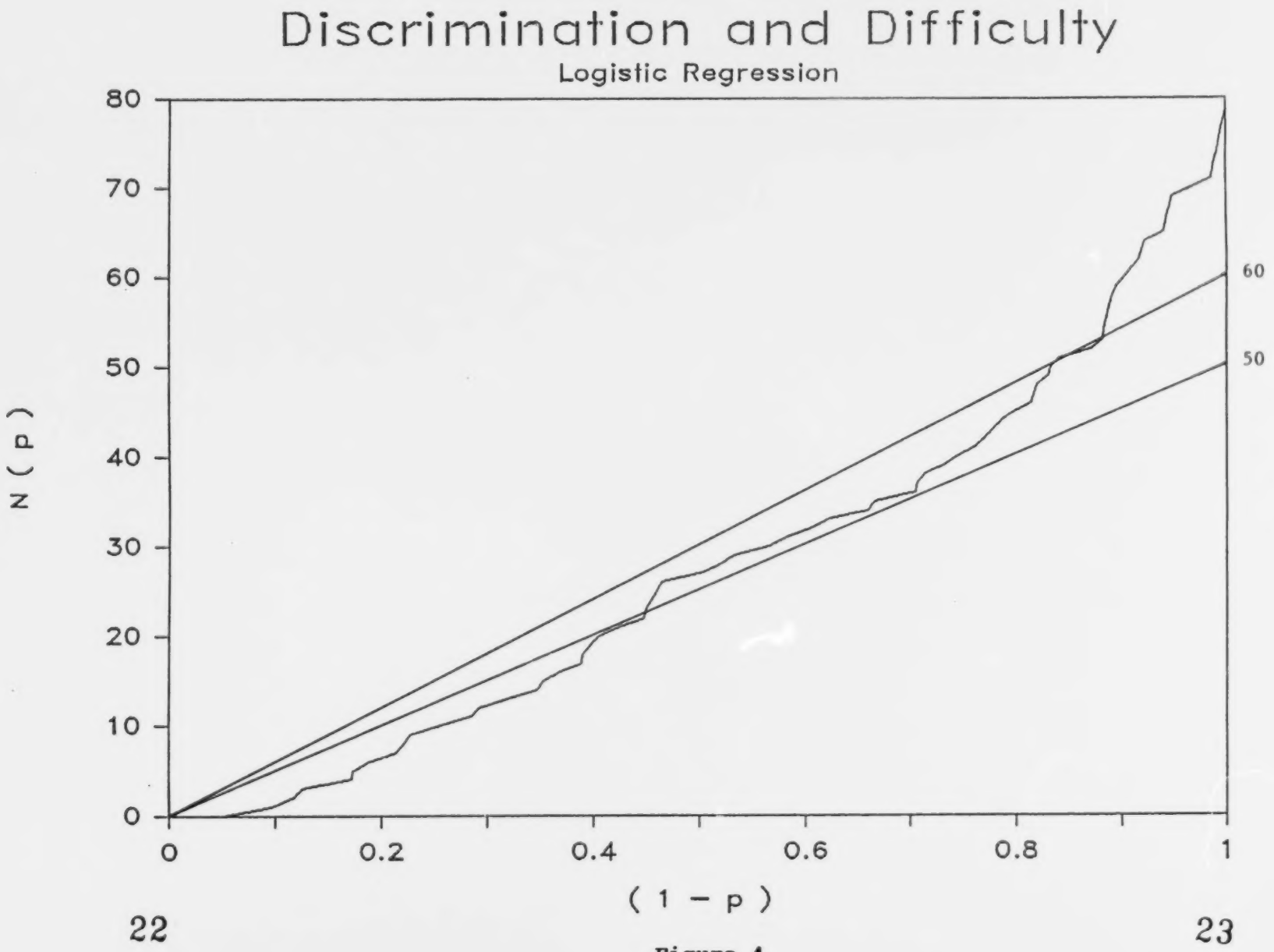

Discrimination and Difficulty
Logistic Regression
80
70
60
50
40
30
20
10
0
N (p)
0
0.2
0.4
0.6
0.8
1
(1 − p)
60
50
22
23

Figure 4

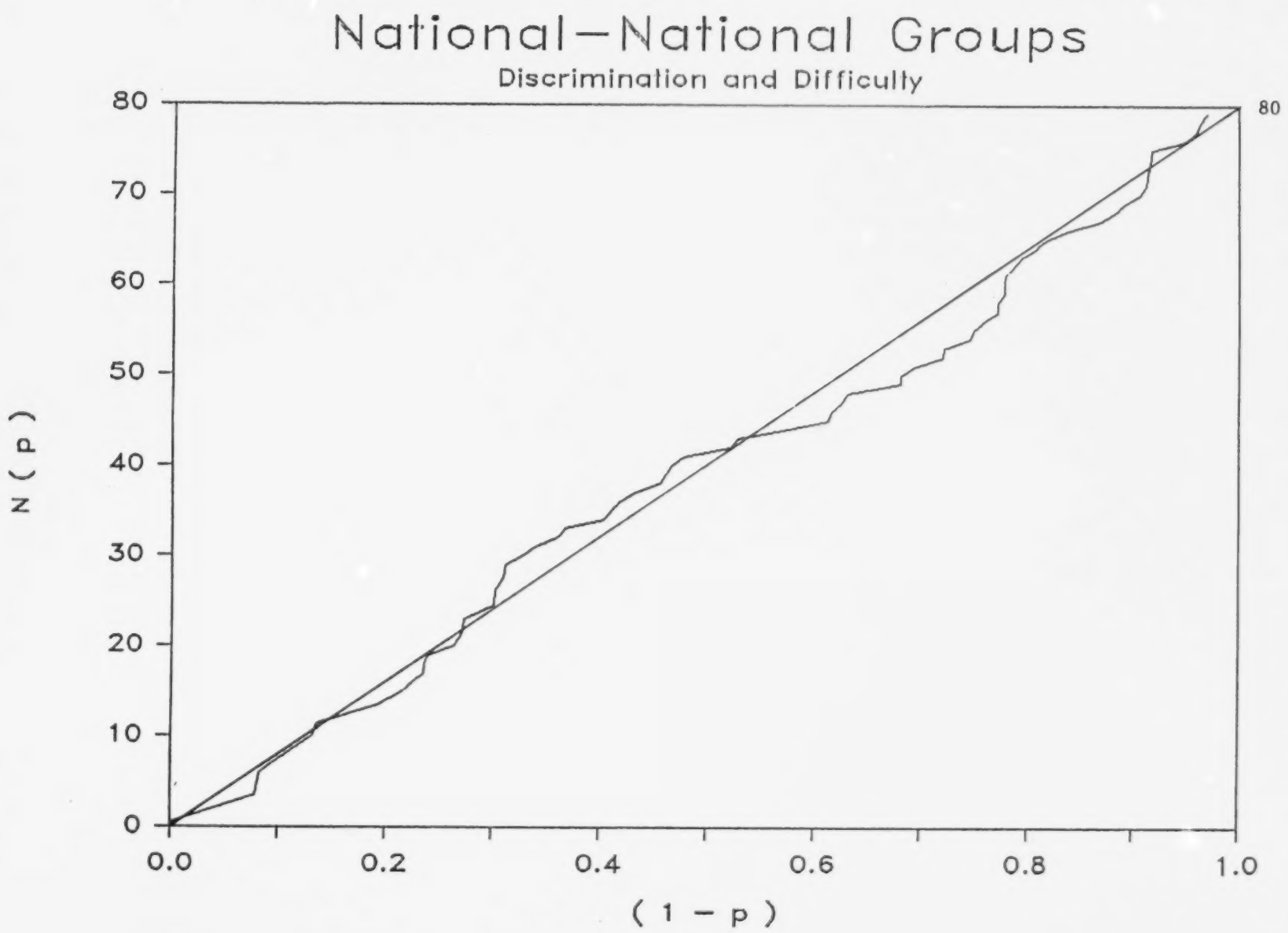

National—National Groups
Discrimination and Difficulty
80
70
60
50
40
30
20
10
0
N (p)
0.0
0.2
0.4
0.6
0.8
1.0
80
(1 − p)

Figure 5

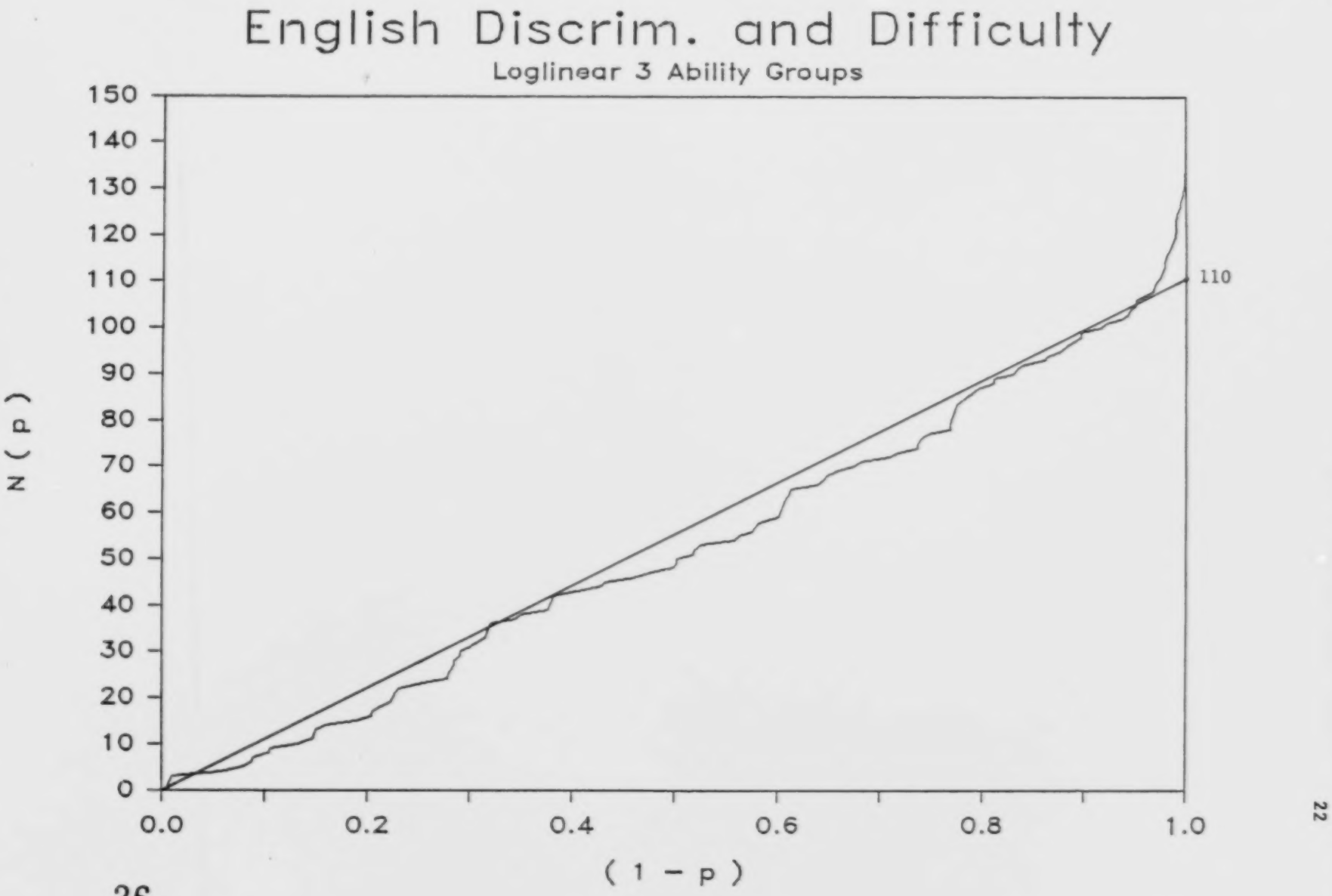

English Discrim. and Difficulty
Loglinear 3 Ability Groups
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
N (p)
110
0.0
0.2
0.4
0.6
0.8
1.0
(1 - p)

Figure 6

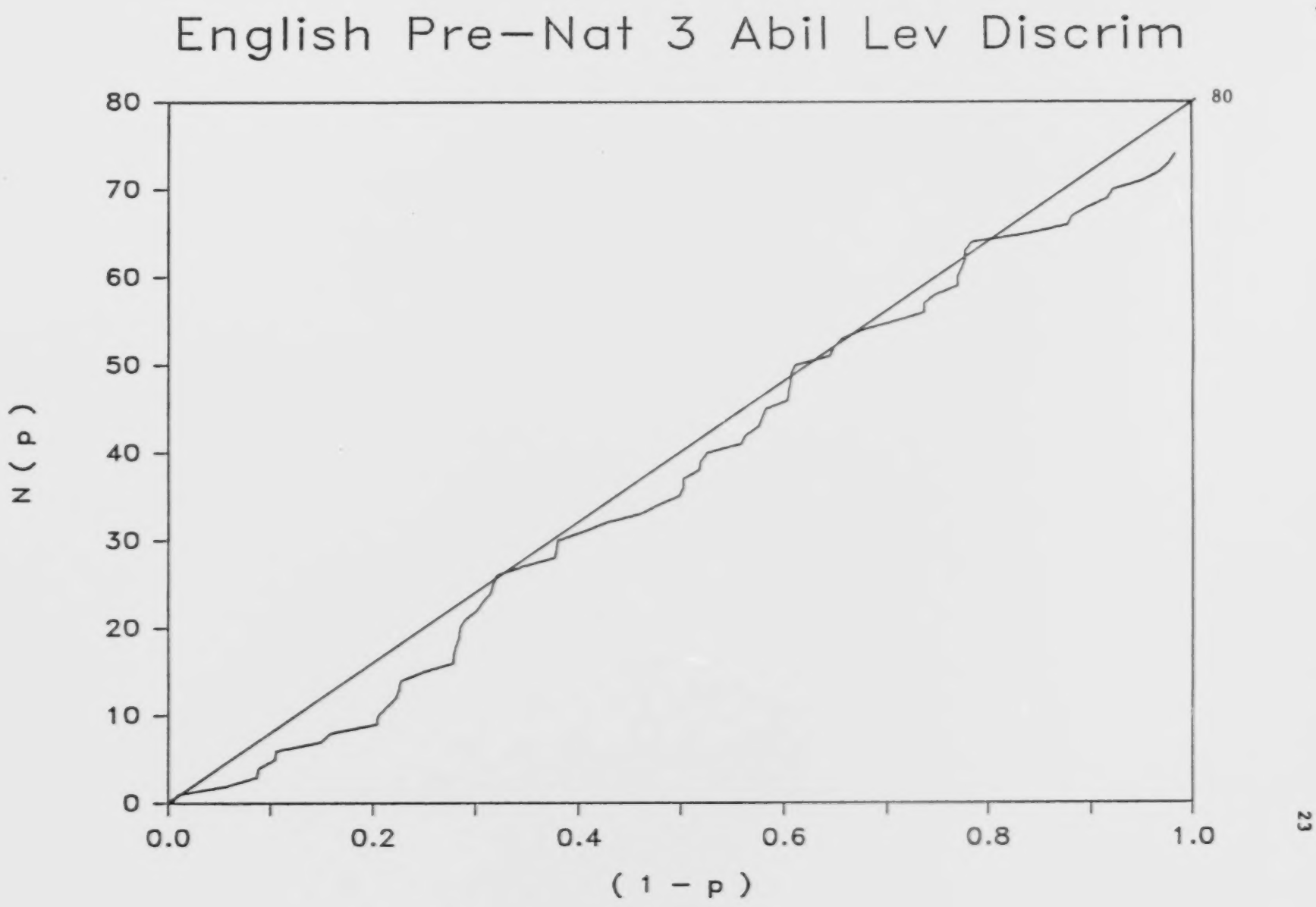

English Pre-Nat 3 Abil Lev Discrim
80
70
60
50
40
30
20
10
0
N (p)
0.0
0.2
0.4
0.6
0.8
1.0
(1 - p)
80

Figure 7

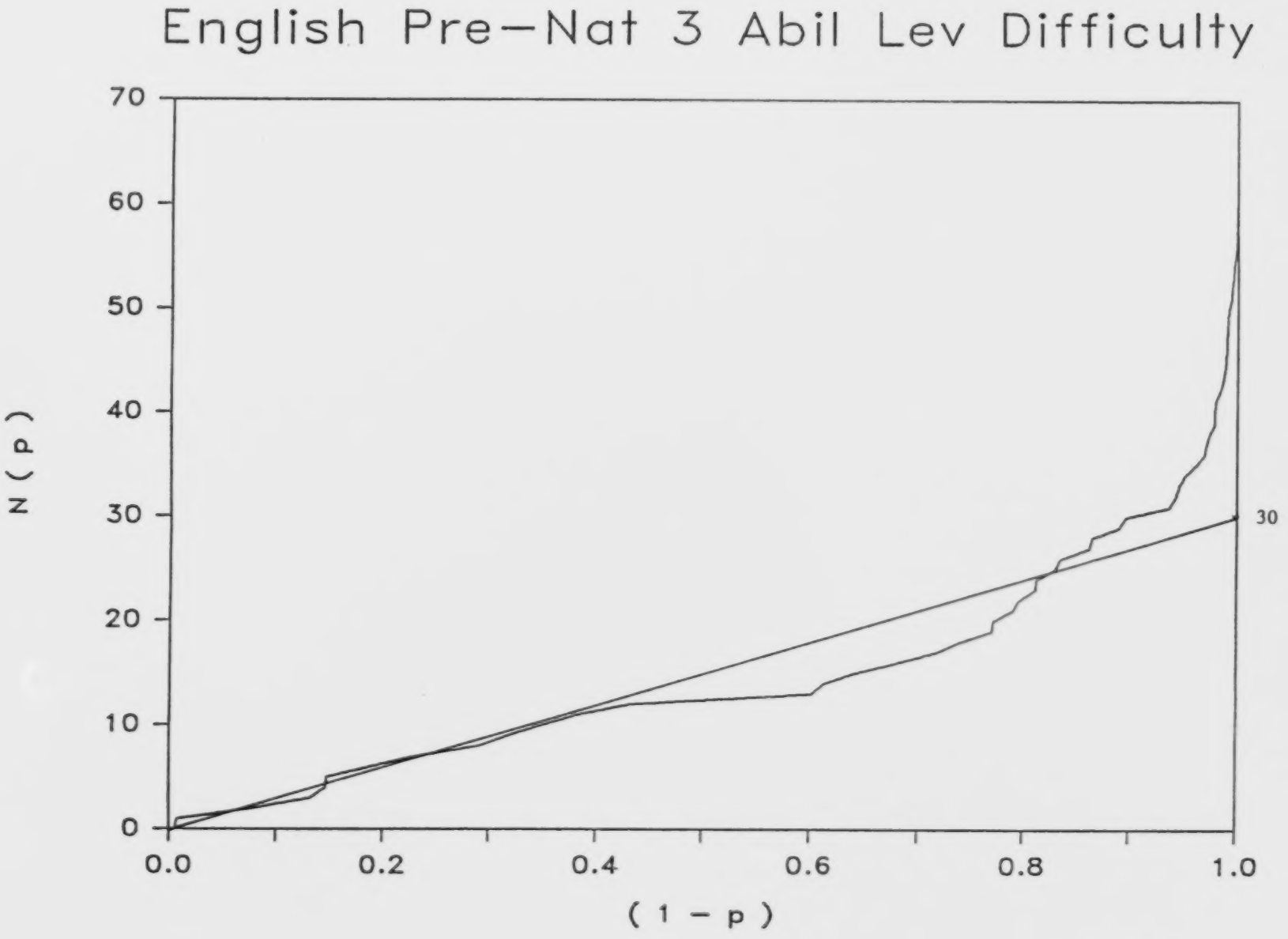

English Pre—Nat 3 Abil Lev Difficulty
70
60
50
40
30
20
10
0
N (p)
30
0.0
0.2
0.4
0.6
0.8
1.0
(1 – p)

Figure 8

END

U.S. DEPT. OF EDUCATION

OFFICE OF EDUCATIONAL
RESEARCH AND
IMPROVEMENT (OERI)

ERIC

DATE FILMED

OCT.14.1986